DOCUMENT RESUME

ED 190 664 TM 800 463

AUTHOR Conway, Malcolm J.
TITLE The Selection and Evaluation of Tests. Director's Handbook: Topics in Testing, Measurement, and Evaluation.
INSTITUTION New Jersey State Dept. of Education, Trenton.
PUB DATE 79
NOTE 16p.: Colored pages may not reproduce clearly.
AVAILABLE FROM State Department of Education, Division of Operations, Research and Evaluation, 225 W. State St., Trenton, NJ 08625 (free)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Advisory Committees: Criterion Referenced Tests: Elementary Secondary Education: *Evaluation Criteria: Information Sources: Models: Needs Assessment: School Districts: *Standardized Tests: Testing Programs: Test Reliability: *Test Selection: Test Validity

ABSTRACT

An advisory committee representing counselors, curriculum specialists, students, teachers, and community members should participate in selecting standardized tests for school districts. These steps should be followed: (1) identify test purpose in relation tc needs and objectives--diagnosis, placement, career guidance, program evaluation: (2) describe group to be tested, consider costs, and predetermine minimum acceptable levels of reliability and validity--a reliability coefficient of .80 and a validity coefficient of .50 are generally suggested cutoffs: (3) check sources of test reviews--Mental Measurements Yearbook and the companion Tests in Print, publishers' manuals, journals, textbooks, and consultants: (4) narrow the search: (5) evaluate the remaining tests against predetermined criteria: (6) conduct a pilot test: (7) submit recommendations to the school board: and (8) monitor testing program implementation. Four indicators of quality for criterion-referenced tests are suggested. (A needs assessment checklist, list of objectives, and test evaluation checklist are appended.) (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190664

DIRECTOR'S HANDBOOK TOPICS IN TESTING MEASUREMENT AND EVALUATION

RECEIVED ERIC/TM

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
M. Conway

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800463

NEW JERSEY STATE
DEPARTMENT OF EDUCATION
DIVISION OF OPERATIONS,
RESEARCH AND EVALUATION
Fred G. Burke
Commissioner of Education

THE SELECTION AND EVALUATION OF TESTS

Malcolm J. Conway

Introduction

This section of the Handbook deals with two separate but related issues. First is the process for selecting a published test to serve a particular purpose from the vast array of published school tests. Second is the task of evaluating the test according to some set of criteria before the final decision is made to use the test. Before discussing how tests are selected and evaluated, it is appropriate to answer the question, "Who should be involved in the process of selecting and evaluating tests?"

A Model for Broad Based Involvement in the Test Selection Process

This question can be answered from both a legal and practical standpoint. In the context of the T & E process, the assessment phase (including identification of standards and "means of measurement" such as tests) is carried out by "the chief school administrator in consultation with the teaching staff." However in the evaluation phase of T & E, as in the goal-setting phase, involvement of community members is mandated, with good reason. Opinions and comments made by a variety of individuals affected by the testing program - students, parents, community members, counselors and curriculum specialists - are useful and desirable.

In large districts, district testing advisory committees of upwards of ten to twenty members are commonplace. The membership is usually composed of guidance and counseling staff, representatives from child study teams, curriculum specialists, students, and teachers. This committee is typically chaired by the Directors of Testing or Pupil Personnel Services. Additional testing committees may operate at the building or departmental level for specific programs, e.g. the selection of tests for particular subject matter areas. Smaller school systems may have only a single committee, under the leadership of the principal, superintendent or guidance director which serves all the schools.

In situations where it is logistically or otherwise impossible to have representative input in the initial phases of the test selection process, the Test Director may want, as an alternative, to identify the instruments and use some type of review process for obtaining the reaction of others. For example, in a single administrator district, the Guidance Director could initially identify potential tests and later involve teachers, guidance and counseling staff,

students and community members to review the recommendations and make final decisions as to which tests to use. In this case particularly, the Test Director may also use outside consultants, if available, to help provide technical assistance in the process.

Bollenbacker (1976) has identified three features common to all testing committees and suggests appropriate activities for the committees, regardless of their organizations. They are: 1) appointment of the committee; 2) charge to the committee; and 3) the actual work of the committee.

Formation of the committee can take place in a variety of ways. Nominations can be made, volunteers can be solicited, direct appointment can be used, or a combination of techniques. A chairperson/facilitator who possesses the necessary qualifications and experience to provide technical assistance and assume a leadership role is essential. Other appropriate roles include resource people to identify tests in specific areas; cataloguers/compilers to gather the necessary copies of specimen tests and related information; outside consultants such as evaluation consultants, physchometricians and psychologists with specific expertise in testing and measurement.

The committee charge should be clearly spelled out. The charge to the committee should include the scope of their task.

1. Will their work include recommendations for teacher-made tests or be limited to published tests?

2. Will program evaluation be included in the committee's purview or will their tasks be limited to making recommendations for test selection alone?

The duties and responsibilities of the committee and its members should also be identified. Resources which will be made available to the committee, time lines, and final products should also be identified. The relationship of the testing committee to other district committees should be specified. It is also essential to provide committee members with copies of all existing policy statements and written documents relating to testing and evaluation and apprise them of any other conditions which may affect their work.

Once the initial organizational tasks of the testing committee have been completed, the members are faced with the process of test selection and evaluation. Figure 1 shows a general model for the process consisting of eight major steps. While the model could be applied to the process of selecting/developing classroom (teacher-made) tests, it is primarily applicable to published, standarized tests. Standarized tests can be defined as commercially prepared tests that provide methods of obtaining information under uniform conditions. The seven steps in this process are detailed in the following sections.

A General Model for TEst Selection
and the Evaluation of Standardized Tests
for a School Testing Program

Identify PURPOSE (objectives) or functions for the given testing program.

Develop TEST SELECTION AND EVALUATION CRITERIA.

Consult SOURCE OF INFORMATION on instruments and obtain specimen copies for review.

NARROW THE SEARCH to a reasonable number of alternative instruments.

EVALUATE the final alternative instruments against pre-stated criteria.

Conduct PILOT TEST of recommended instrument(s).

Prepare report of process and SUBMIT RECOMMENDATIONS for adoption by school board/superintendent; and,

MONITOR testing program implementation as well as participate in on-going review ofpprocess.

Purpose

In deciding which test(s) to select, the first step should be to systematically determine the purposes of testing. In some districts formal written documents describing purposes for testing may already exist. These might be a board policy pertaining to testing, a philosophy statement on testing or rationale for a testing program, or general goal statements for the testing program on more specific objectives. In the absence of detailed specifications for testing, a committee might be formed to gather this information by conducting a needs assessment of the current testing program.

Appendix A Shows an instrument that may be used to obtain needs assessment information. A related Appendix B shows a sample rationale for a hypothetical district testing program and associated objectives.

If the committee must refine an existing program, it should gather data through surveys, interviews or

similar techniques from the users of the existing program. This investigation will help to identify areas of the testing program in need of improvement.

Regardless of the techniques used to determine the purposes for testing, typically three justifiable uses of test results are commonly identified: classroom, guidance, and administrative. Classroom functions include identifying students in need of remediation, determining achievement levels for individual students, and determining growth in achievement of individuals or groups of students. Examples of guidance functions include measurement of students' attitude toward the world of work and helping students set career and vocational goals appropriate to their aptitudes. Administrative functions include appropriate placement of new students in the curriculum, curriculum (program) evaluation, curriculum planning, evaluation of materials, and providing information (public relations).

Once the purposes for and use of testing have been determined, the committee must link these to the specific objectives (content) of the subject matter to be tested. Mehrens and Lehman (1969) point out that in addition to the problem of matching the test and subject matter content, there is also a problem in determining exactly why you want to use the test. Their admonition in this regard is worth keeping in mind when deciding to use a standardized test in favor of the many alternatives, e.g. contracts, teacher-made tests, conferences. They state:

"It cannot be emphasized too strongly that the first and most important steps in test selection are to determine exactly why you are giving the test, what type of information you expect from it, and how you intend to use it".

Test Selection and Evaluation Criteria

General Characteristics:

While technical adequacy (reliability, validity, etc.) is the *sine qua non* for the selection of any test, a number of general considerations can be used to narrow the potential pool of tests to a reasonable number for more detailed evaluation. In addition to the aforementioned purpose for testing, several other characteristics can be mentioned. Brown (1971) recommends that the test selector obtain general information about the group tested, their age, special abilities, socioeconomic background, intelligence, etc. He also suggests attention be paid to the practical considerations of time, cost (both the direct costs of material and scoring and indirect costs of personnel time for preparation, administration, and interpretation) and administrative provisions including special training, facilities, or equipment. Mehrens and Lehman (1969) suggest the added general characteristic of format--attractive appearance, appropriate type size, quality illustrations.

Technical Characteristics

Reliability:

The physical measurements which we make (e.g. height, weight, speed) generally yield consistent results and hence we can place confidence in our measurements.[2] This is due in large part to the stability of physical traits.

In contrast, such a high degree of consistency is not possible in psychological measurement due to the fluctuation in human behavior from time to time and situation to situation. Test scores are influenced by a variety of extraneous factors that produce inconsistency or error. For example, an examinee may be more fatigued or anxious today than tomorrow.

If we are to have confidence in our tests, test scores and the subsequent decisions we make, we need to be concerned with the problem of consistency of measurement from two standpoints: 1) of determining how consistent our measures are and 2) of improving the consistency through minimizing error. Reliability is the general term used to describe consistency of measurement. Reliability is usually experienced in terms of a coefficient ranging in value from 0 to +1.0. A reliability coefficient of 1.0 indicates perfect (error-free) measurement or consistency. Three major types of reliability are stability, equivalence and homogeneity.

The first method of estimating reliability is to determine how stable the scores are on successive administrations of a test. In the test-retest method used to determine stability, the test is given to a group and re-administered to the same group after a delay varying from hours to a week or more. Scores from the first administration are then plotted against

2. It should be noted that even in physical measurement there is some inconsistency or unreliability. For example, an error may occur in measuring weight on a bathroom scale on successive days. The recorded weights usually differ because the scale is inaccurate or because the person doing the measuring may have read the numbers incorrectly.

scores from the second administration and a correlation coefficient is computed. Although there are no hard and fast rules for achievement tests, correlation coefficients (test-retest) of .90 or above are desirable. Correlations between .59 and .80 are generally questionable, while correlation coefficients below .50 are unacceptable.

A second method of estimating reliability is appropriate when you are using different forms (alternate forms) of the same instrument; for example, in a pretest-post-test situation with form 1 used for the pretest and form 2 for the post-test. If different forms of the test are equivalent or parallel (measuring the same skills, knowledge, or attitudes) scores on the two forms should be comparable. In the parallel form technique, one form is given to a group and the second form is administered to the same group at some later time. Scores from the first form are then correlated with scores from the second or parallel form. Here again reliability coefficients of .80 or above are desirable.

A third method of estimating reliability is to determine if the test is internally consistent; that is to what extent do the test items assess the same skill. This type of reliability is often referred to as homogeneity.

Two methods for establishing internal consistency are used. The first is called the "split-half" method. It involves a single test administration after which half the items are scored for a half-test score and the other half of the items are scored similarly. Scores from the two halves are corrected and the resultant coefficient is corrected (by the Spearman-Brown Formula) to compensate for the fact that the actual test was twice as long as the two halves that were scored to compute the correlation coefficient. Another method to determine internal consistency uses the Kuder-Richardson Formula 20 (coefficient alpha). This coefficient represents the average score from computing all possible combinations of split-half reliabilities. Practically speaking, this type of reliability is most critical in a test which is comprised of several subtests each purporting to measure a separate skill or concept.

Generally, internal correcting reliability coefficients are higher (for a given test) than test-retest coefficients.

Thus, it can be seen that a test does not have a single reliability coefficient. Rather, a test has different coefficients for different uses. Strictly speaking, reliability is not an inherent characteristic of the test. It is a function of scores on the test obtained by the group. Therefore, the similarity between the group on which reliability data was gathered and the group for which the test will be used is important to consider.

Validity:

The American Psychological Association's Standards for Educational and Psychological Tests and Manuals (1973) is an essential reference for anyone considering selecting or evaluating tests. It defines validity information as "the degree to which the test is capable of achieving certain aims". That is, validity indicates the degree to which certain judgments or inferences can be made on the basis of the test items. This judgment may take the form of scores on the test predicting or estimating an individual's future standing or the degree to which the individual possesses a hypothetical trait or quality measured by the test. From our brief discussion it should be apparent that a valid test is necessarily a reliable one; unless a test yields consistent results it cannot be said to be measuring what it purports.

Three types of validity information are usually identified: 1. content; 2. construct; and 3. criterion-related (predictive or concurrent). Content validity judges whether or not the test contains a representative sample of tasks which define the population of tasks or test items of interest (content domain). While it is fairly simple to specify the content domain for certain content areas (e.g. elementary math), most content domains are so extensive that full enumeration is almost impossible. It is helpful to list curricular or instructional objectives of a school program in order to determine if the sample of test items does indeed match the curriculum content. As Martuzza (1977) points out, "Unless informed judgment can be made about whether a particular item is contained in the content domain of interest, the determination of content validity is impossible." The process for establishing content validity includes first defining or describing the domain and then asking experts to judge whether the test items are a representative sample of the domain. Melville proposes a simple point-value scale for judging the relevance of each test item under consideration and admonishes

the selector to ask of each item: "What particular skill, understanding, type of material or subject matter does this item seem to be covering?" A recent NIE study (1977) has uncovered what have been termed "striking differences" in content among various standardized tests, providing further impetus for careful analysis of the content of each item and its validity for the decisions at hand.

Construct validity is evaluated by examining the degree to which certain psychological constructs-e.g. "anxiety", "reading readiness", "mechanical aptitude" account for performance on a test. For example, construct validity could be established experimentally for a test purporting to measure anxiety by trying out the test on people whom experts say are anxious. If the "high anxiety" people obtain high scores on the instrument, then the instrument can be said to have construct validity. Correlational methods are usually used.

Criterion-related validity involves determining the relationship between an individuals score on the test and his most probable standing on some other variable termed a criterion. If the prior test performance is used to estimate subsequent performance, the criterion-related validity is termed "predictive validity". Alternately, concurrent validity indicates the relationship between scores on the present test and present standing on the criterion, e.g. another test or different evaluative measure.

Like reliability, there are no fixed rules for the magnitude of the validity coefficient. Generally, validity coefficients are lower than reliability coefficients since validity involves two distinct measurements of the same content. Thus, coefficients in excess of .75 are generally thought of as high. Coefficients in the range of .50 - .75 are acceptable, while those below .50 are usually unacceptable.

Sources of Information

The prospective test selection has many sources of information available about educational and psychological tests. Major sources include reference books, test publishers, journals, text books, and experts.

Reference Books: Two major reference books edited by Oscar K. Buros are often used as a starting point for test selection. They are the annual Tests in Print, a comprehensive bibliography of educational, psychological, and industrial tests, and the companion, Mental Measurements Yearbook, published every several years, with the eighth edition now in preparation. In addition to basic information on the test--title, author, publication date, scores, forms, levels, etc.--the Yearbook also contains a list of published articles and research on the test. The Center for the Study of Evaluation at UCLA has also published a comprehensive evaluation of published instruments in a series of books, which includes higher order cognitive, affective, and interpersonal tests, preschool, and kindergarten tests. A unique feature of this series is the assignment of letter "grades" by reviewers for each of several rating system characteristics, e.g., examine appropriateness. Still other reference books are compendia of tests in specific areas. Johnson and Boomarito's Tests and Measures in Child Development (1971) is an example.

Publishers: Test publishers will usually provide catalogues of their tests to qualified persons. The catalogues typically contain descriptions of the tests, costs, purpose, grade levels, norms, scoring methods, and costs of specimen sets. Publishers also periodically issue bulletins at little or no cost dealing with a variety of specific topics related to tests and testing. The specimen copies of tests available from publishers are among the most useful resources permitting the selector to judge the actual content, format, and technical quantity.

Journals: Other sources of information are educational and psychological journals. Some journals publish actual test reviews, others report studies using tests, still others treat the psychometric aspects of testing. Examples of journals are The Journal of Educational Measurement (quarterly), American Educational Research Bulletin (quarterly), Review of Educational Research (five times a year), Educational and Psychological Measurement (quarterly), Psychological Bulletin (monthly), and Measurement and Evaluation in Guidance (quarterly).

Text Books: Numerous textbooks on educational and psychological measurement typically include chapters containing descriptions and evaluation of specific tests or major categories of tests. Although not as comprehensive as the reviews, texts such as Aiken (1976), Anastasi (1968), Brown (1971), Cronbach (1960), Lindvall (1975), and Thorndike and Hagen (1969), often provide more detail for specific tests concerning the purpose and construction.

Experts: A fifth source of information on tests is the expert. Large city school districts, most colleges and universities, and educational resource centers, e.g., EICs, have staff members with responsibilities and expertise in testing who can answer specific testing questions or route users to appropriate people or agencies.

Narrow the Search

To select a limited number of instruments meeting the general criteria of appropriateness for the group to be tested, practical considerations of cost and administrator's time, etc., Brown (1971) suggests beginning with Buros' publications and then proceeding to journals and textbooks.

If no available test meets the general requirements, the criteria may need to be changed; a locally developed test may be constructed to meet the criteria, or another means of measurement may need to be used.

Evaluate Test Against Criteria

After the initial narrowing down of the potential instruments a more detailed evaluation of the selected instruments can be made in terms of the criteria that have been previously established.

Several excellent alternative formats for evaluating tests have been proposed, Brown (pps. 241-250), Cronbach (p. 187) and Kirby, et al (pps. 68-70), among others, provide detailed forms that can be used as checklists. They differ primarily in the level of specificity of the questions they ask and the information which is sought. A general format is shown in Appendix C.

Pilot Test

Where feasible the recommended test should be administered to a sample of students under conditions as close to the actual administration as possible, e.g. physical setting, time, instructions. Participants' reaction to the test and the testing situation should be gathered to uncover any unanticipated problems.

Submit Recommendations and Monitor Program

Once the pilot test has been completed, the committee should document the entire process used in evaluating and selecting specific alternative instruments. This information will assist the superintendent and/or board in making appropriate decisions.

The committee should continue to function as monitors of the implementation of the testing programs, per se. This may entail gathering teacher perceptions of the usefulness of the information the test provides.

Special Considerations in Evaluating a Criterion-Referenced Test

In comparison to the literature on norm-referenced tests, there are relatively few published evaluations at present of commercially-available criterion-referenced tests although an increasing number of reviews have appeared in research and curriculum-related journals. It is anticipated that the Ninth edition of Buros Mental Measurements Yearbook will help fill this void. The CSE Test Evaluation Kits, (Hoepfner, 1970) do contain evaluations of certain criterion-referenced instruments, e.g. the Prescriptive Reading Inventory, IOX Criterion-Referenced Tests. Fitzgibbon and Morris (1975) provide one of the few recommendations on this subject. They suggest selectors examine four indicators of test quality.

Reliability

Do all of the items which are to measure any one objective seem to be measuring the same skill?

They should, of course. For example, a test measuring ability to multiply improper fractions might contain a series of multiplication items consisting of improper fractions. However, no proper fractions should be included since this involves a different level of skill. The formats should also be consistent.

How many items per objective does the test include?

A minimum of four or five items is recommended for individual mastery determination.

Validity

Do the items test the objective they claim to?

Each item should be examined to insure that it measures behaviors that provide sufficient evidence of mastery.

Is the test format appropriate for the age/ability of examinees?

This indicator is difficult to judge in tests for young children who have difficulty with complex directions or scoring systems.

Content Analysis

This judgmental process involves matching the test's objectives with the objectives of the program.

Format Analysis

The essential question asked is "Do the item

formats on the test compare to those that are learned by students?" If, for example, the student who has learned subtraction in vertical format is provided with a test item in horizontal format, the inconsistency in format can prevent the student from displaying competency in a skill he or she has mastered.

Martuza (1977) provides an excellent discussion of appropriate techniques for determining the reliability and validity of criterion-referenced tests. For example, the analagous process to the norm-referenced test-retest procedure for determining reliability involves determining the ability of the test to classify students consistently as masters or non-masters on two separate test administrations. This can simply take the form of the percentage (proportion) of examinees classified in the same category--"master" or "non-master" on successive test administrations.

APPENDIX A
Sample Testing Program Needs Assessment
(Adapted from Instrument Developed by: B. Baylor, CTB/McGraw Hill)

You are being asked to assist in the needs assessment of the present testing program of your school district.

The purpose of the needs assessment is to identify those objectives of a testing program that will enable the teachers, specialists, counselors, and administrators to make more effective use of test information.

The following contains statements which describe the kinds of information that a comprehensive testing program can provide. Please respond to the testing objectives you feel are appropriate to your role in the following two ways:

IMPORTANCE — How important do you feel the appropriate testing objective is in a comprehensive testing program for your school and/or district? In the space provided for the evaluation of each of the objectives on the left, place a 3, 2, or 1 according to the importance that you feel the objective has in a testing program. Use the values below to rate the appropriate objectives.

3 Very Important
2 Somewhat Important
1 Of Little Importance or Not Needed

PRESENT — How adequately are the testing objectives appropriate to your role being attained at the present time in your school and/or district? Please respond to this question for the applicable objectives in the space provided for on the right of the paper. Use the values below to rate how adequately the applicable objectives are presently being attained.

3 Adequately
2 Partially
1 Not At All

Please make sure that you evaluate only the objectives you feel are appropriate to your role in each of the two ways.

A TESTING PROGRAM NEEDS ASSESSMENT

The following statements describe the kinds of information that a comprehensive testing program can provide. Please respond only to the objectives you feel are appropriate to your role in each of two ways, 1. the importance of the objective and 2. the degree to which the objective is presently being reached.

YOUR ROLE (check 1) _____ teachers ____specialist ____counselor ____ administrator

IMPORTANCE		PRESENT
_____	1. To determine the level of each student's ability for academic work.	_____
_____	2. To evaluate the achievement level of each student in relation to his ability.	_____
_____	3. To diagnose the instructional needs of each student.	_____
_____	4. To identify the number of students in a classroom with specific skills and concepts not yet mastered.	_____
_____	5. To prescribe remedial instruction where needed, for individual students.	_____
_____	6. To determine needs in school's curriculum in order to make decisions relative to curriculum revision.	_____
_____	7. To maintain an accurate record of each student's academic growth from year to year.	_____
_____	8. To determine the average achievement levels attained by groups of students are commensurate with the average student's capabilities.	_____
_____	9. To facilitate grouping for instructional purposes within a classroom according to skills deficiencies.	_____
_____	10. To determine if instructional objectives are being attained in each school.	_____
_____	11. To compare, each student's achievement level to the level obtained by a normative group.	_____
_____	12. To compare the average achievement of groups of students to that obtained by a normative group.	_____
_____	13. To compare for administrators the achievement levels among schools within a district or among groups of schools.	_____
_____	14. To determine student's aptitudes for various careers and occupations.	_____
_____	15. To measure a student's attitude toward the world of work.	_____
_____	16. To measure a student's motivation for learning.	_____
_____	17. To help select appropriate classroom materials and techniques in instruction.	_____
_____	18. To develop awareness of consistency or change in the performance of groups of students over time.	_____

Counselors appropriately use test results as an aid in:

1. Helping pupils develop realistic self-expectations and set appropriate educational and career goals.
2. Identifying those children for whom special learning provisions may be desired.
3. Improving understanding of individual problem situations and improving communication with other school personnel and parents about such problems.
4. Preparing information to guide discussions with parents about their child.
5. Making plans for the placement of incoming individuals and pupil groups.

Administrators within each school building appropriately use test results as an aid in:

1. Evaluating the effectiveness of curricular provisions in meeting certain major objectives of the school.
2. Interpreting educational successes and needs of the local school to parents and to the community.
3. Identifying special provisions required to meet the unique needs of individual pupils and groups.
4. Providing information needed by appropriate outside individuals and institutions, such as potential employers, other schools, colleges, and accrediting agencies.
5. Working with other administrators in identifying instructional needs common to all schools.

Administrators and specialists in the central office appropriately use test results as an aid in:

1. Assessing the effectiveness of the total school program in meeting certain major objectives of the system.
2. Interpreting educational needs and successes of the system to the community at large.
3. Analyzing instructional needs for the purposes of determining special provisions or requesting special funding for individual schools or the system.
4. Conducting project evaluation and research studies.

APPENDIX B

Sample Rationale and Objectives for District-wide Testing Program

Rationale:*Adapted from Milwaukee, Wisconsin Public Schools document, 1971.

The overall purpose of the measurement program is to establish a data base useful for the improvement of instruction. Data provided are expected to facilitate decision-making at all levels--by the board, by administrators and specialists in the central office and in the schools, by teachers, by parents, and by pupils.

Objectives:

Sufficiently mature pupils appropriately use test results as an aid in:

1. Understanding themselves as they are.
2. Identifying academic areas of strength for purposes of further growth and progress.
3. Identifying academic areas of weaknesses for purposes of improvement.
4. Developing reasonable expectation and setting realistic educational and vocational goals.

Teachers appropriately use test results as an aid in:

1. Identifying special needs of pupils to be considered in daily instruction.
2. Grouping pupils for special learning situations within a class.
3. Identifying pupils who have special needs for diagnosis and planning.

4. Determining present and expected achievement levels for individual pupils and classroom groups.
5. Identifying needs for instructional emphasis in certain curricular areas.
6. Communicating individual pupil academic skill development and learning needs to the pupil and the parent.
7. Increasing pupil and parent interest in educational achievement.
8. Selecting appropriate classroom materials and techniques of instruction.
9. Developing awareness of consistency or change in the performance of individuals and groups over time.

APPENDIX C
General Checklist for Evaluating a Test

Directions:

A, B, and C should be used, as appropriate, to rate the individual characteristics. In column A the respective weights are established and listed for each characteristic prior to using the checklist. A simple 3 point scale-- 1 = slightly important; 2 = moderately important; 3 = highly important-- can be used. A similar scale can be used in column B to assign a rating of the degree to which the characteristic is present for the test under consideration-- 1 = to a slight degree; 2 =to a moderate degree; 3 = to a great degree. In column C the product of the pre-determined weight and assigned rating is entered as a "score" for each characteristic. The form can be adapted, as appropriate to local situations.

CHARACTERISTIC	A IMPORTANCE (WEIGHT) 1 = Slightly 2 = Moderately 3 = Highly	B PRESENT 1 = To slight degree 2 = To a moderate degree 3 = To a great degree	C SCORE (A X B)
I. General Features			
A. Name			
B. Publication Date			
C. Author/Publisher			
D. Forms and Levels			
E. Related Materials, (manuals, scoring, guides, etc.)			
F. Norm or criterion-referenced			
II. Purposes			
A. Author's/Publisher's			
B. Your purposes			
C. Other reported uses			
III. Practical Features			
A. Time			
1. Total administration time			

CHARACTERISTIC	A IMPORTANCE (WEIGHT) 1=Slightly 2=Moderately 3=Highly	B PRESENT 1=To slight degree 2=To a moderate degree 3=To a great degree	C SCORE (A X B)
III. Practical Features (Continued)			
2. Scoring time			
B. Cost			
1. Test booklets (reusable?)			
2. Answer sheets			
3. Machine scoring			
4. Related materials			
5. Staff time to administer/procter			
C. Format			
1. Quality (printing, illustrations, layout)			
2. Answer sheets appropriate for age of students and their ability?			
3. Directions clear? (administration, scoring, responding, interpreting)			
D. Items			
1. Total number of items and item coverage by subtests			
2. Appropriate readability			
3. Appropriate format			
4. Do items have face validity?			
5. Evidence of item review to eliminate sexism/other biases?			
6. Items matched to content to indicate relative emphasis?			

CHARACTERISTIC	A IMPORTANCE (WEIGHT) 1=Slightly 2=Moderately 3=Highly	B PRESENT 1=To a slight degree 2=To a moderate degree 3=To a great degree	C SCORE (A X B)
IV. Technical Features			
A. Validity			
1. Evidence?			
2. Types			
3. Generalizability to other contexts?			
B. Reliability			
1. Evidence?			
2. Types			
3. Reliability of subtest?			
4. Reliability for other groups?			
C. Norm			
1. Types			
2. Quality (accuracy, sample size, composition)?			
3. Local norms if appropriate			
D. Empirical Studies/ reviews			
1. Evidence of empirical studies			
2. Quality as reviewed			
3. Particular Strengths			
4. Any limitations?			
V. Overall Rating			

About the Author . . . .

Malcolm J. Conway serves as Coordinator of Research and Evaluation for the Educational Improvement Center - Central. A certified biology teacher and administrator having university teaching experience, he received his undergraduate degree from Tufts University. He also holds a masters degree in Educational Research and Evaluation from the University of Massachusetts where he is completing his Ed.D.

Prior to his current position he was a staff member of ERIE B.O.C.E.S. in Buffalo, New York. His particular professional interests include pupil and program evaluation, criterion-referenced testing, and instructional applications of computers. He has served as a consultant in the design and operation of testing programs to a variety of districts in New Jersey and other state

REFERENCES

Aiken, L.R. Psychological and Educational Testing. (2nd ed.) Boston: Allyn and Bacon, Inc., 1976.

American Psychological Association. Standards for Psychological Tests and Manuals. Washington,D.C. Author, 1973.

Anastasi, A. Psychological Testing. (3rd ed.), New York: MacMillan Co., 1968.

Bauernfeind, R. Building a School Testing Program. Bensenville, Illinois: Scholastic Testing Service, 1968.

Bollenbacker, J.K. Selecting and Using an Advisory Committee. Guide for School Testing Programs. East Lansing, Michigan: National Council on Measurement in Education, 1976.

Brown, Frederick G. Measurement and Evaluation. (1st ed.) Illinois: F.E. Peacock Publishers, Inc., 1971.

Buros, O.K. The Eighth Mental Measurements Yearbook. New Brunswick, New Jersey: Rutgers University Press, 1976.

Cronbach, L.J. Essentials of Psychological Testing. (2nd ed.) New York: Harper and Row, 1960.

Fitz-Gibbon, C.J. & Morris, L.L. How to Measure Achievement. Los Angeles, California: Center for the Study of Evaluation, UCLA, 1975.

Goslin, D.A., Epstein, R.E. & Halloci, B.A. The Use of Standardized Tests in Elementary Schools. New York: Russell Sage Foundation, 1965.

Gronlund, N.J. Preparing Criterian Referenced Tests for Classroom Instruction. New York: The MacMillian Co., 1973.

Hoepfner, R., Stern, C., & Nummendal, S.G., (Eds.) CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and the Early Childhood Research Center, UCLA, 1971.

Hoepfner, R., Stern, C., & Nummendal, S.G., (Eds.) CSE Tests of Higher Order Cognitive, Affective, and Skill. Los Angeles: Center for the Study of Evaluation, UCLA, 1971.

Johnson, O.G. & Bonmarito, J.W. Tests and Measurements in Child Development: A Handbook. San Francisco: Jossey-Bass Inc., 1971.

Kansas State Department of Education. A Position Paper on Standardized Testing. Unpublished manuscript, Author, 1972.

Kurley, J.H., Culp, W.H., and Kurley, J.: MUST-Manual for Users of Standardized Tests. Bensenville, Illinois: Scholastic Testing Service, 1973.

Lindvall, C.M. & Nitko, A.J. Measuring Pupil Achievement and Aptitude. New York: Harcourt, Brace, Javanovich, 1975.

Martuza, V.R. Applying Norm and Criterian-Referenced Measurement in Education. New York: Allyn and Bacon, 1977.

McKenna, B. & Quinto, F. Alternatives to Standardized Testing: A Handbook. Washington, D.C.: National Education Association of the United States, 1977.

Mehrens, W.A., & Lehman, I.J. Standardized Tests in Education. New York: Holt, Rene rt and Winston, 1969.

Melville, S.D. Selecting an Achievement Test. Princeton, New Jersey: Educational Testing Service, Undated Memo.

NIE Council Debates Standardized Testing — While Research Shows It's Not At All Standard. Education U.S.A., 1977.

Thorndike, R.L., and Hagen, E. Measurement and Evaluation in Psychology and Education. (3rd ed.) New York: John Wiley and Sons, Inc., 1969.

Wick, John W. Educational Measurement: Where Are We Going and How Will We Know When We Get There? Ohio: Charles E. Merrill Publishing Co., 1973.

PUBLICATIONS IN THIS SERIES INCLUDE:

Volumn I.	Purpose of Testing	*David W. Pearsall*
Volumn II.	Test Selection and Evaluation of Tests	*Malcolm J. Conway*
Volumn III.	Organizing and Managing a District Testing Program	*Robert L. Jackson, II*
Volumn IV.	Some Uses of Standardized Test Results	*Eugene J. Travers*
Volumn V.	Motivation and Test Wiseness	*David F. Engiehardt*

Published by the New Jersey Department of Education, 225 West State Street, Trenton, New Jersey 08625. The opinions expressed in this publicaition are those of the invited authors and do not necessarily reflect the position or policy of the New Jersey Department of Education.

ACKNOWLEDGMENTS

Thanks are extended to the Planning Committee for this publication and to the editorial review panel consisting of Dr. Curtis Banks, Princeton University; Dr. Douglas Penfield, Rutgers University; and, Dr. Paul Van R. Miller, Lehigh University.

Appreciation is also expressed to Ms. Vivian Sobke for preparation of the manuscripts.